ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Models, Algorithms and Data (MAD): Introduction to Computing

Spring semester 2018

Prof. P. Koumoutsakos, Prof. J. H. Walther
ETH Zentrum, CLT
CH-8092 Zürich

Mockup Exam

Issued: June 11, 2018
Hand in: June 11, 2018

Exam details.

This is a mockup exam.
It is meant to give you an idea of the
style of questions that you may expect in the exam.

We are not planning to provide solutions to these problems. The exam will cover all topics from the lecture (blackboard and lecture notes) and from the exercises. Topics that will **not** be part of the exam in 2018: Fourier Transforms, Markov Chain Monte Carlo, and Rejection Sampling.

There will be three main types of exercises:

1. *Theory*. This part covers overall knowledge about applicability and properties of all methods and concepts introduced in the lecture and exercises.
   This part corresponds to approximately 40% of the total points in the exam.

2. *Numerical problems*. This part covers the mathematical and numerical concepts introduced in the lecture and exercises. For instance, you might be asked to perform the following tasks (on paper):

   - apply an appropriate numerical method to a given problem
   - prove some property of a given numerical method
   - modify/adapt a given numerical method for a different setup
   - ...

   This part corresponds to approximately 50% of the total points in the exam.

3. *Pseudo-codes*. This part is a pseudo-code writing problem for some specific numerical methods.
   This part corresponds to approximately 10% of the total points in the exam.

The point distribution mentioned here is approximate, and might vary in the final exam.

Exam directives. In order to pass the exam, the following requirements have to be met:

- Read carefully the first two pages of the exam. Write your name and Legi-ID where requested. Before handing in the exam, **PUT YOUR SIGNATURE ON PAGE 2**.

- Clear your desk (no cell phones, cameras, etc.): on your desk you should have your Legi, your pen, your notes and optionally your non-scientific calculator.

- The teaching assistants will give you the exam sheets and the necessary paper sheets. You are not allowed to use any other paper sheets. On the top-right corner of every page write your complete name and Legi-ID.

- The personal summary consists of no more than 4 pages (2 sheets). The personal summary can be handwritten or machine-typed. In case it is machine-typed, the text has to be single-spaced and the font size has to be at least 8 pts. You are not allowed to bring a copy of somebody else's summary.

- You can answer in English or in German; the answers should be handwritten and clearly readable, written in blue or black - do NOT write anything in red or green. Only one answer per question is accepted. Invalid answers should be clearly crossed out.

- If something is disturbing you during the exam, or it is preventing you from peacefully solving the exam, please report it immediately to an assistant. Later notifications will not be accepted.

- You must hand in: the exam cover, the sheets with the exam questions and your solutions. The exam cannot be accepted if the cover sheet or the question sheets are not handed back.

Family Name:	
Name:	
Legi-ID:	

Question	Maximum score	Score	TA 1	TA 2
1	8			
2	6			
2	8			
3	6			
4	14			
5	16			
6	18			
7	16			
8	12			
9	14			
10	22			
11	16			
Total	156			

With your signature you confirm that you have read the exam directives; you solved the exam without any unauthorized help and you wrote your answers following the outlined directives.

Signature:

Theory

Question 1: Lagrange vs Cubic (8 points)

Compare pros and cons between interpolation using Lagrange and Cubic Splines. Write two features that Lagrange is better than Cubic Splines and vice versa.

Question 2: Least Squares on a circle (6 points)

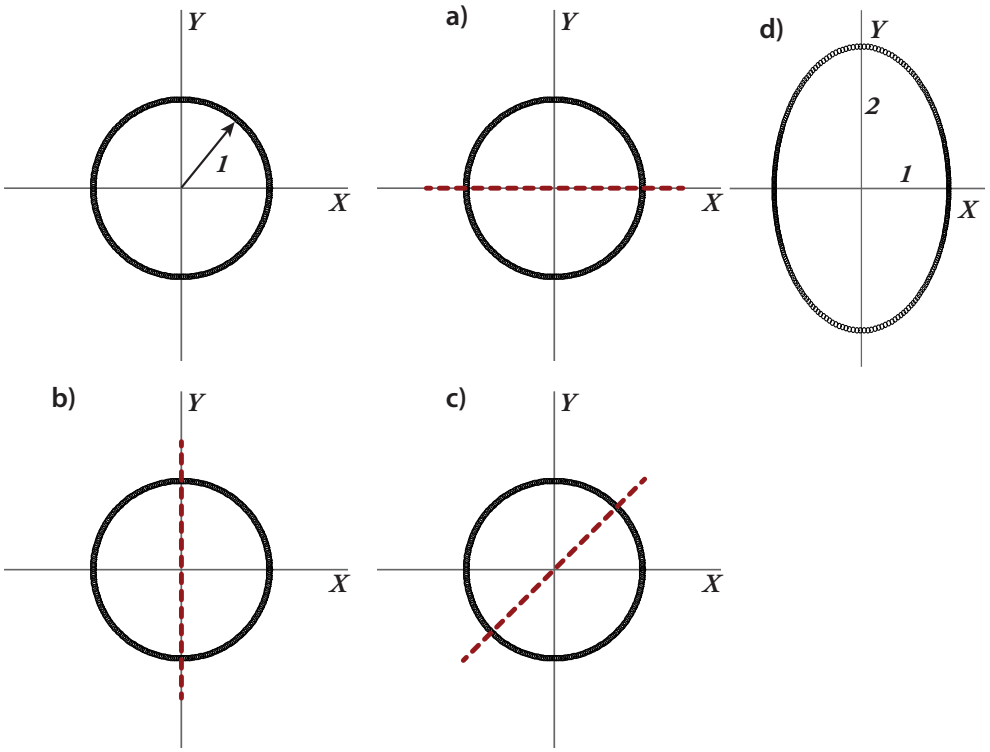

Figure 1: Data points (small black dots) are located on a circle (top left panel). In a) - c), dashed line depicts possible LSQ fittings $y = ax$ of this data. In d) - data on an ellipse.

The data points in a very big dataset ($N \gg 1$) are uniformly (and densely) located on a unit circle with its center in origin (see Fig. 1).

a) What is the correct straight line obtained with LSQ method for $y = ax$ on this data? Choose one variant and explain your answer:

- Horizontal line through origin, Fig. 1 a)
- Vertical line through origin, Fig. 1 b)
- Diagonal line through origin, Fig. 1 c)
- Any line through origin is correct
- Other variant, please elaborate.

Hint: $a^2 + (b-a)^2 \geq \frac{1}{2}b^2$.

b) How will the answer change if the data is located on a vertically stretched ellipse (see Fig. 1, d)?

Question 3: Orthonormal Functions (8 points)

Assume that you are given a set of basis functions $\{\psi_1(x), \psi_2(x)\}$ on the interval $[-1, 1]$ which are orthonormal with respect to the inner product $< \cdot, \cdot >$. We approximated some function $f(x)$ on the interval $[-1, 1]$ using $\psi_1(x)$ and $\psi_2(x)$ as $f(x) \approx \alpha_1\psi_1(x) + \alpha_2\psi_2(x)$.

Answer the following questions:

1) You wish to add a new basis function $\psi_3(x)$. Write a formula to make this function orthonormal with respect to the two others.
2) You wish now to approximate f by the set $\{\psi_1(x), \psi_2(x), \psi_3(x)\}$ as

$$f(x) \approx \tilde{\alpha}_1\psi_1(x) + \tilde{\alpha}_2\psi_2(x) + \tilde{\alpha}_3\psi_3(x).$$

   What is the relation between α_1 and $\tilde{\alpha}_1$, α_2 and $\tilde{\alpha}_2$?

Question 4: Radial Basis Functions (6 points)

Choose an appropriate answer:

1) An advantage of using Radial Basis Functions (RBF) is that one doesn't need to recompute previous coefficients when adding a new basis function. (yes / no)
2) Centers of RBF are always located at the data points. (yes / no)
3) RBF approximation problem can always be represented as a linear system of equations. (yes / no)

Question 5: Overfitting vs underfitting (14 points)

When trying to find the "correct complexity" of a model one often faces problems of overfitting to the data vs underfitting.

a) Briefly explain the difference between these two problems and why the problems happen.

b) On the two pictures below (Figure 2) draw how underfitting could look like on the left plot and how overfitting could look like on the right plot for the given data set.

c) In case of overfitting do you expect a small or a big prediction error for a new data point?

d) Are there methods to avoid overfitting? Name one.

Question 6: Trapezoidal rule (16 points)

a) Mark the following statements as true or false.
   1. In the trapezoidal rule there are points that are used in two different subintervals. Therefore we must evaluate the function we want to integrate twice at every point.
   2. The formula for the trapezoidal rule in an interval $[a, b]$ is given by the equation:

$$I = \int_a^b f(x)dx \approx \frac{h}{2}\left[f(a) + f(b) + \sum_{j=1}^{n-1} f_j\right] \tag{1}$$

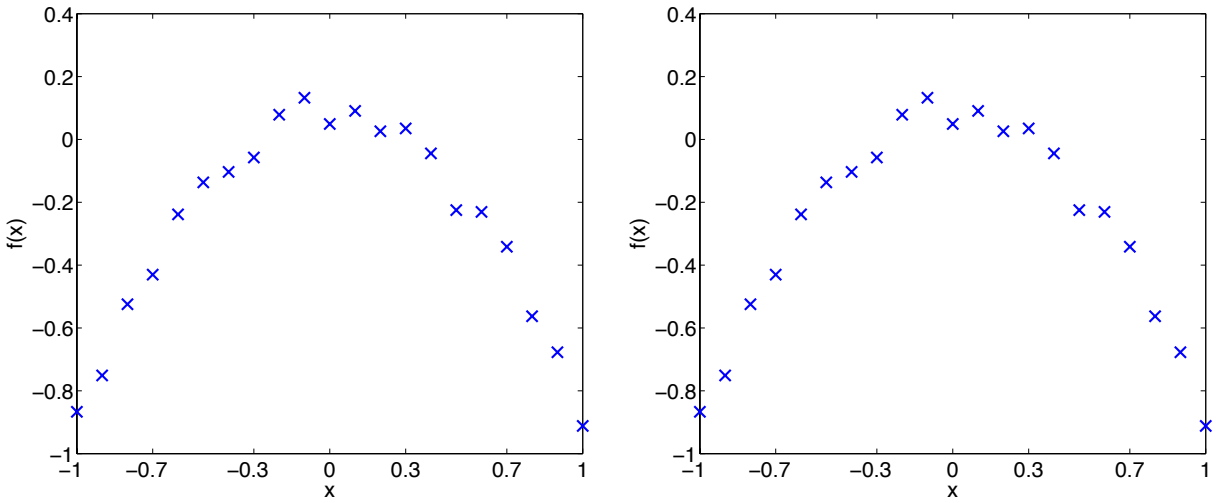

Figure 2: Overfitting vs underfitting.

3. When h is halved in the trapezoidal rule, half of the function values used with step length h/2 are the same as those used for step length h.
4. The trapezoidal rule with two sub-intervals is exact for integrating at most second order polynomials.
5. The rectangle and the trapezoidal rule have the same order of accuracy.

b) You are interested in buying a very modern apartment in Zurich. The owner has given you a plan of the apartment (see figure). Estimate the area of the entire apartment (*including* areas below furniture and kitchen appliances) in m^2 *using the Trapezoidal rule*. Explicitly show all the intermediate steps you have performed, including the choice of quadrature points. Is your estimate exact?

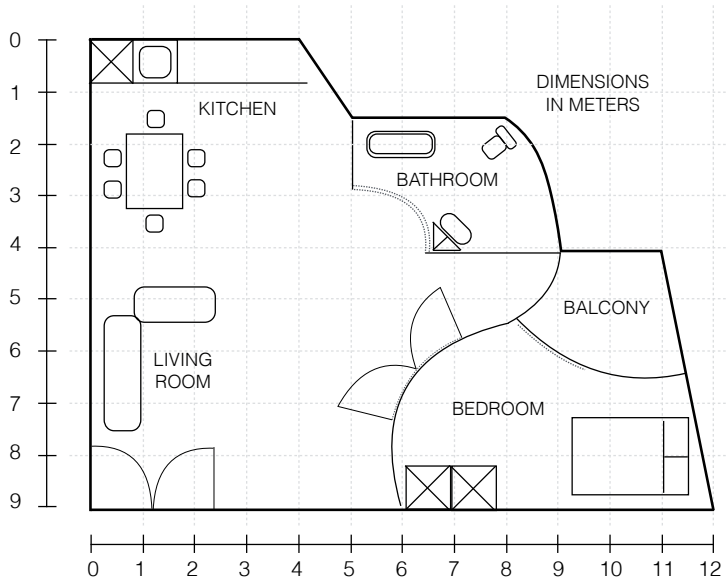

Figure 3: Figure for Q2: Top plan of the apartment of interest.

Numerical Problems

Question 7: Least Squares method (18 points)

You are preparing software for automatic processing of specific scientific data. The data you receive from the experimentalists is a set of points (x_i, y_i) for $i = 1, 2, \ldots, N$ with $N \geq 3$. This data is often perfectly linear, namely $y_i = \alpha x_i + \beta$. But unfortunately errors are inevitable in experiments, and sometimes several points (called "*outliers*") appear to deviate from the straight line: if j-th point is an outlier, then $y_j = \alpha x_j + \beta + e_j$. We will assume that parameter α is known, and you are planning to use Linear Least Squares (LSQ) to identify parameter β of the linear dependence, but the outliers may introduce errors in the fit. Therefore you want to try and minimize the bad influence of the outliers.

a) What method (algorithm) studied in the course can be used to automatically identify the outliers? Explain how.

For the following questions assume there is only one outlier (x_j, y_j) in the dataset.

b) Given α, β, dataset (x_i, y_i), with $y_i = \alpha x_i + \beta$ for $i = 1, 2, \ldots, N$ if $i \neq j$, and an outlier $y_j = \alpha x_j + \beta + e_j$, what is the coefficient β^{LSQ} of the linear fit computed with the Linear Least Squares method? In this special case, the answer should be simplified to depend only on the true value β, the outlier error e_j and the total number of points N.

c) You have an idea to change the LSQ method to the Least Absolute Residuals (LAR). Similarly to LSQ, LAR method can be written as a minimization problem (note that explicit solution for LAR cannot be easily derived analytically, unlike LSQ),

$$\beta^{LAR} = \arg\min_{\hat{\beta}} \sum_{i=1}^{N} |(\alpha x_i + \hat{\beta}) - y_i|. \tag{2}$$

With the same assumptions as in the previous question, what is the coefficient β^{LAR} of the linear fit computed with the Linear Absolute Residuals method? Again, the answer should be simplified to depend only on the true value β, the outlier error e_j and the total number of points N.
Hint: $\frac{d|c|}{dc} = \frac{c}{|c|}$; for simplicity, you can assume that $e_j > 0$ and that $\beta \leq \beta^{LAR} \leq \beta + e_j$.

d) Compare the exact β with estimated β^{LSQ} and β^{LAR}. Which method is better to deal with outliers?

Question 8: Boundary conditions for cubic splines (16 points)

When constructing a cubic spline from a given set of data points (x_i, y_i), the system of equations that must be solved can be written as:

$$A_i f''_{i-1} + B_i f''_i + C_i f''_{i+1} = D_i \tag{3}$$

where

$$A_i = \frac{\Delta_{i-1}}{6} \tag{4}$$

$$B_i = \left(\frac{\Delta_{i-1} + \Delta_i}{3}\right) \tag{5}$$

$$C_i = \frac{\Delta_i}{6} \tag{6}$$

$$D_i = \frac{y_{i+1} - y_i}{\Delta_i} - \frac{y_i - y_{i-1}}{\Delta_{i-1}} \tag{7}$$

and

$$\Delta_i = x_{i+1} - x_i. \tag{8}$$

You decide that for your boundary conditions, you'll assume constant curvature in the two intervals closest to the end points (i.e., $f_1'' = f_2''$, and $f_N'' = f_{N-1}''$). Write down the matrix system that you'd need to solve for constructing the desired spline. Make sure to explicitly write out at least the first three and the last three rows of the system, in terms of $A_1, B_1, \cdots$ and any other variables you deem necessary.

Question 9: Interpolation and Extrapolation (12 points)

You are the number one trader in Silverman Sachs, and your boss often consults you for opinions. Today, your boss comes to you with a set of revenue data from a new startup company. He is interested in predicting the performance of this company in the future.

Figure 4 shows the complete data set $D = \{(t_i, R_i)|i = 1, ..., 9\}$.

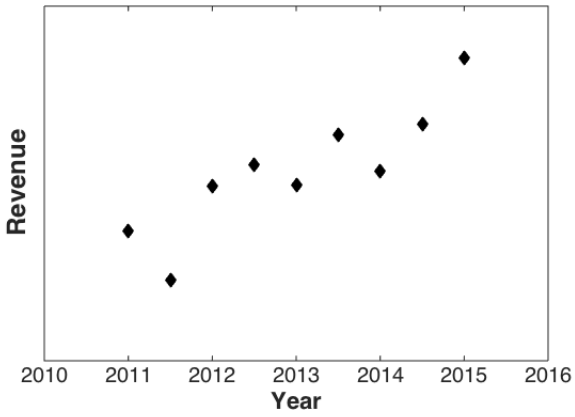

Figure 4: Revenue data for a new startup company.

a) As a first trial, your boss wants you to use Lagrange Interpolation for the prediction. Write down the expression to evaluate the revenue $\tilde{R}$ given time $\tilde{t}$ using Lagrange Interpolation. Your answer shall be written in terms of t_i and R_i for $i = 1, \ldots, 9$. (Tips: use summation or product notation for simpler form of answer.)

b) Your boss wants to estimate the revenue of this company on March, 2013 using the Lagrange Interpolation. He asks you to comment on the performance of this estimation. Do you expect the estimation to be accurate? Why?

c) To show off your knowledge base, you introduce two more interpolation methods to your boss:

1. Cubic Splines
2. A set of radial basis functions F_i for $i = 1, \ldots, 9$ with the following form,

$$F_i = \begin{cases} f(|t - t_i|), & \text{if } |t - t_i| < 2\,\text{years} \\ 0, & \text{otherwise,} \end{cases}$$

where t_i are from the data set D, and $f(\,\cdot\,)$ is some arbitrary function.

Do you expect the estimation for March 2013 using these two methods to be better or worse than the estimation using Lagrange Interpolation? Why?

d) Your boss wants you to predict the revenue of the company in 2018. What would you expect for the performance of the three methods mentioned above? (i.e., Can you foresee what will be the prediction value from each method? If yes, what is the value and why? If no, why not?)

Question 10: Newton Cotes formulas (14 points)

a) Use the Newton-Cotes formulas for $n = 1$ to compute the coefficients

$$C_k^n = \frac{1}{b-a} \int_a^b l_k^n(x) dx, \quad k = 0, \ldots, n, \tag{9}$$

where $l_k^n(x)$ are Lagrange polynomials in interval $[a, b]$ of degree n:

$$l_k^n(x) = \frac{(x - x_0)(x - x_1) \ldots (x - x_{k-1})(x - x_{k+1}) \ldots (x - x_n)}{(x_k - x_0)(x_k - x_1) \ldots (x_k - x_{k-1})(x_k - x_{k+1}) \ldots (x_k - x_n)}, \tag{10}$$

and where x_i are equidistant points in $[a, b]$. For $n = 1$ that is: $x_0 = a,\ x_1 = b$.

b) Using the computed coefficients C_k^n from (9), derive the resulting numerical integration rule using the Newton-Cotes formula

$$I \approx (b - a) \sum_{k=0}^{n} C_k^n f(x_k). \tag{11}$$

c) How is this integration rule called? Which order polynomials are integrated exactly using this rule? Which order accurate is this rule (just state the order, no need to prove)?

Question 11: Romberg integration with Simpson's rule (22 points)

The Simpson's integration rule reads as follows,

$$I = \int_a^b f(x) dx \approx I_S = \frac{b-a}{6} \left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right), \tag{12}$$

and is 5-th order accurate in a given interval $[a, b]$, i.e.

$$I_S - I = \mathcal{O}(h^5).$$

a) Compute the error of the composite Simpson's integration on an interval $[0, 1]$ divided into N subintervals.

b) Similarly to a classical Romberg method, we define I_0^n as an approximation of I with the Simpson's rule using n intervals. Moreover, I_k^n is a higher-order approximation obtained by Richardson extrapolation using I_{k-1}^{2n} and I_{k-1}^n. Using the error expression obtained in previous subquestion, derive a formula for I_k^n which defines the Romberg integration with underlying Simpson's rule, for any arbitrary n and k.

Pseudo Codes

Question 12: Adaptive Quadrature using Trapezoidal rule (16 points)

In the class, you learned that an adaptive scheme can be used to improve efficiency of a numerical scheme. The basic idea is that you only refine subintervals which have error higher than your demanded threshold. This can avoid wasting computational power in subintervals that already reach sufficient accuracy. One important issue in this approach is to estimate the error in a subinterval based on a chosen numerical scheme. In practice, calculation of the exact error is not feasible. In this question, you will investigate on optimizing the error estimation method for the trapezoidal rule.

a) For a given subinterval $\{x_i, x_{i+1}\}$, the numerical integration of $I_i = \int_{x_i}^{x_{i+1}} f(x)\,\mathrm{d}x$ using the trapezoidal rule, denoted as I_{T_i}, can be written as

$$I_{T_i} = I_i + \frac{1}{12} f''(x_{i+1/2}) h^3 + O(h^5),$$

where $h = x_{i+1} - x_i$ and $x_{i+1/2}$ is the midpoint between x_i and x_{i+1}, i.e., $x_{i+1/2} = (x_i + x_{i+1})/2$. Assuming the full integral $I = \int_a^b f(x)\,\mathrm{d}x$ is subdivided into N equally spaced subintervals with length h, show that the total error for approximating I using trapezoidal rule on each subinterval is in the order of h^2, i.e.,

$$I_T = \sum_{i=0}^{N-1} I_{T_i} = I + O(h^2).$$

b) Following Richardson's idea for error estimation, let the target quantity $G = I$, and the approximation method $G(h) = I_T(h)$, the trapezoidal rule as a function of subinterval length h. Since the trapezoidal rule is a higher order approximation scheme, the expression for error estimation shown in the notes, $\epsilon(h/2) \approx G(h/2) - G(h)$, is an overestimation of the actual error. Please improve the error estimation based on the expression $I_T = \sum_{i=0}^{N-1} I_{T_i} = I + O(h^2)$.

c) Write a pseudo-code for a subroutine called *AdaptiveTrapez* that estimates a 1D integral $I = \int_a^b f(x)\,\mathrm{d}x$ by applying adaptive Trapezoidal rule for given tolerance ε. The input of *AdaptiveTrapez* includes the boundaries a and b and the tolerance ε. The output of *AdaptiveTrapez* is a single scalar estimate of the integral I. In *AdaptiveTrapez*, you may call a function evaluation subroutine *F(x)* that takes in one scalar input x and return a

scalar that is the function value evaluated at x. You may write any necessary additional subroutines and use them inside *AdaptiveTrapez*.

Good luck!